Medicine Radar – A tool for exploring online health discussions

Krista Lagus[1], Minna Ruckenstein[2], Atte Juvonen[3], and Chang Rajani[3]

[1] Methods Centre, Faculty of Social Sciences, University of Helsinki, Finland
krista.lagus@helsinki.fi
[2] Consumer Society Research Centre, Faculty of Social Sciences, University of Helsinki, Finland
minna.ruckenstein@helsinki.fi
[3] Futurice Oy

Abstract. Research focusing on online health discussions provides valuable insights into the use of medicines, as well as health-related experiences and difficulties currently not well understood. We introduce Medicine Radar, a tool for exploring health-related online discussions obtained from the Finnish Suomi24 chat forum. The health subset of the entire Suomi24 data consists of 19 million messages written over a time span of 16 years. We outline the method, identify some challenges in analyzing Finnish texts and explain how we overcame them in this specific domain. In particular, we present a novel method for generating domain vocabularies from colloquial texts, which utilizes a combination of machine learning and human input. Medicine Radar is accessible as an open sourced web interface that we hope will inspire and facilitate further research.

Keywords: social media, lightly supervised machine learning, medical texts, vocabulary discovery, open data, open source

1 Introduction

Medicine Radar (Lääketutka) is a tool for exploring health-related online discussions in the Suomi24 dataset. Suomi24 is a popular anonymous message board in Finland and the dataset contains tens of millions of messages from a time span of 16 years. The opening of the data, its properties and possibilities for analysis have been described in [11]. Medicine Radar was created as an open collaboration between researchers of Citizen Mindscapes research consortium and Futurice's Chilicorn Fund, aiming to produce a workable digital tool for social-scientifically oriented social media research. We provide this tool as a web interface which is open source, and designed to be accessible and usable to the general public.

Health and its absence is one of the most common topics in Suomi24. Anonymous discussions online are known to provide an important arena for addressing and sharing health-related problems [3]. In particular, medicines are in many ways central to the way in which health and illness are discussed [17]. Medicines

and their dosages, side effects and availability are referred to, for example, when discussing the diagnosis and progress of diseases. In addition, the criticism of health care practices is often associated with limited, inadequate or excessive medication of people and overt medicalization of people's lives. Overall, anonymous social media data opens a perspective to peer-to-peer talk on health and medication, bypassing the way in which these issues are talked about in professional settings and as part of doctor-patient relations. From this perspective, talk about medication can be seen as part of the arena where people are conceptualizing and negotiating their health. Medicine talk can also be emotional, suggesting that the analysis of human-drug relationship opens up emotionally-charged perspectives to experiences of health and recovery.

There is a long history of analyzing, visualizing and exploring online discussions with machine learning methods. In the nineties, Self-Organizing Maps were one of the first methods utilized for large-scale analysis of colloquial discussions using machine learning (see for example [11] and [8]). The majority of existing research in natural language processing is still conducted on English texts. Using similar methods on Finnish and other languages with a high degree of inflecting and compounding has proven to be quite challenging. Whereas in English the vocabulary size tends towards a log curve as the size of the material grows, in Finnish it is typical that as the size of the data set grows, the number of word forms continues to grow linearly. This leads to data scarcity, which is a problem for many machine learning methods that rely on bag-of-words representations of data. Moreover, in colloquial texts, rife with misspellings, acronyms and invented words, the usefulness of linguistic analyses is limited. Section 2 explains how we tackled the unique challenges of health-related colloquial Finnish texts. Section 3 describes the web interface and section 4 outlines use cases.

2 Concept discovery applying augmented intelligence

A key problem for any analysis concerning textual content is the identification of the concepts of a language, and how to map them to the word forms in that language. A particularly had problem is how to recognize relevant concepts from informal discussion data. This entails two separate sub-challenges: first, when does an instance of a word represent an interesting concept *at all*, and second, when do two words refer to *the same* concept. We focused on the concepts of central interest in this domain, namely medicines and symptoms. This focus enabled us to combine a number of information sources and approaches in a way that could perhaps be described as an augmented intelligence method for concept-oriented analysis of colloquial health discussions. The general approach is replicable in other domains as well, where one might require an approach that combines several sources of knowledge, including data analysis, linguistic tools and limited human input. An early version of the method has been utilized for studying rhythms of emotions, in order to derive vocabulary for the emotions "fear/worry" and "joy"[12]. There the data-analysis parts of the method were simulated manually and with the use of online data-driven dictionaries.

2.1 Generating concept vocabularies

A good starting point to recognizing words as either medicine or symptom would be a vocabulary of medicines and symptoms. Symptom word lists in Finnish did not exist to our knowledge. As for medicine lists, we did find some, but they were insufficient for our purposes. The same medicine can have many different marketing names, people may refer to it with a nickname, and typos are very common with medicine names. People were also discussing natural remedies and illegal substances as if they were medicines, and we wanted to capture those discussions as well. It became apparent that we had to create our own vocabularies.

Creating a vocabulary totally manually would be very time-consuming, so we created a tool for collecting "theme words" from a large corpus and used it for generating vocabularies for "medicines" and "symptoms". Basically, the tool explores the space of contextually similar word forms in a greedy manner, utilizing human input for final decisions. The efficiency of the joint process depends heavily on the ordering of the words, which needs to be cognitively easy for the user. Our tool works as follows: the user gives a single seed word. The tool suggests "similar" words based on that seed word. User can accept or reject each word. After a while the suggestions deteriorate and the user can jump to the next seed word, which is automatically chosen from previously accepted suggestions. For similarity, we used Word2Vec [13], which is able to find semantically (and syntactically) similar words based on their context in a large corpus. For example, "burana" may be similar to "panacod" because these words are both medicines and therefore they often appear in similar contexts. In addition, the tool considers frequency of a word in corpus, adds words in stemmed form and memorizes rejected words. The vocabularies were also tweaked by adding some words with regex, some manual editing, and by parsing a list of medicine names that was found online.

2.2 Catching different expressions

As we collected our vocabularies we probably found some expressions for all of the most commonly appearing concepts in the discussion data. For example, we may have found "kipu" (pain) and "kivut" (pains), but at this stage we have not yet mapped them to the same concept. Moreover, it is likely that many forms are missing from our vocabulary, such as "kivulias" and "kipeee" (variants of painful). We would like to locate all these different expressions and map them to the concepts that they belong to. Note that we are not trying to catch synonyms, only different surface forms of the same word (lemma). This task turned out to be much more complicated for symptoms than medicines.

The names of medicines are typically foreign, and are not easily confused with other words in Finnish. For example, we can safely assume that all words which begin with "ibuprof" refer to the same medicine concept, "ibuprofeeni". We used simple insights like this to map together different surface forms for the same medicine concepts. However, in the case of symptom words, which typically follow Finnish fonotactics, similar methods were found to lead to many

errors where we would map non-symptom words into symptom concepts, or to combining unrelated concepts. To tackle these issues we arrived at a more complex solution for mapping symptom expressions. We created a new version of our vocabulary, such that each word has three representations: original, stemmed and lemmatized. For lemmatization we used Finnish Dependency Parser [5]. We then go through this new vocabulary and map together words if one word is a substring of another word, or if the Levenshtein distance between the two word forms is exactly one. When applying the vocabulary for identifying instances of symptom concepts from text, we correspondingly use the lemmatized version of the discussion data. There are also various other tweaks and the complete method is available open source[4].

The final vocabulary of concepts includes approximately 1500 medicine concepts and 500 symptom concepts, for which we describe tens of thousands of surface forms. The theme word collector and the collected vocabularies have been published as open source.

2.3 Calculating relations of concepts

We wanted to explore relationships between concepts in the discussion data and for this purpose we needed a metric to measure the strength of a relationship. After experimenting with different metrics, such as tf-idf, we eventually settled on Lift[5], which is intuitively explainable and similar to the *Average precision* measure that is typically used in Information Retrieval for ranking documents in searches. For example, if we wanted to find concepts most strongly related to *headache*, we would calculate lift values from *headache* to every other concept. Let's say that *painkiller* appears in 2% of all Suomi24 posts, but if we only look at the set of posts where *headache* appears, suddenly *painkiller* appears in 10% of posts. In this example the Lift value from *headache* to *painkiller* would be $10\%/2\% = 5$. An intuitive interpretation for this result would be that *headache* in text lifts *painkiller* by 5x.

3 The Medicine Radar interface

The Medicine Radar web interface can be found at https://www.laaketutka.fi. Nearly all of the source code as open source and the remaining parts are available upon request. A user can search with any medicine or symptom name. The resulting view shows concepts which are strongly associated with the search concept in the discussion data. For example, Figure 1 shows related symptoms for search concept "sleeplessness" (results include "sleepiness", "daytime tiredness", "restlessness", "nervousness", "sleep disturbance", "irritability", and "lack of appetite"). By clicking on the circles, the user can read the actual discussions. For example, Figure 2 displays a view into posts which include both concepts "sleeplessness" and "melatonin". Also note in Figure 2 how different instances for these concepts are highlighted. In case of medicine search, the interface also

[4] https://github.com/futurice/health-visualizations

[5] $https://en.wikipedia.org/wiki/Lift\_(data\_mining)$

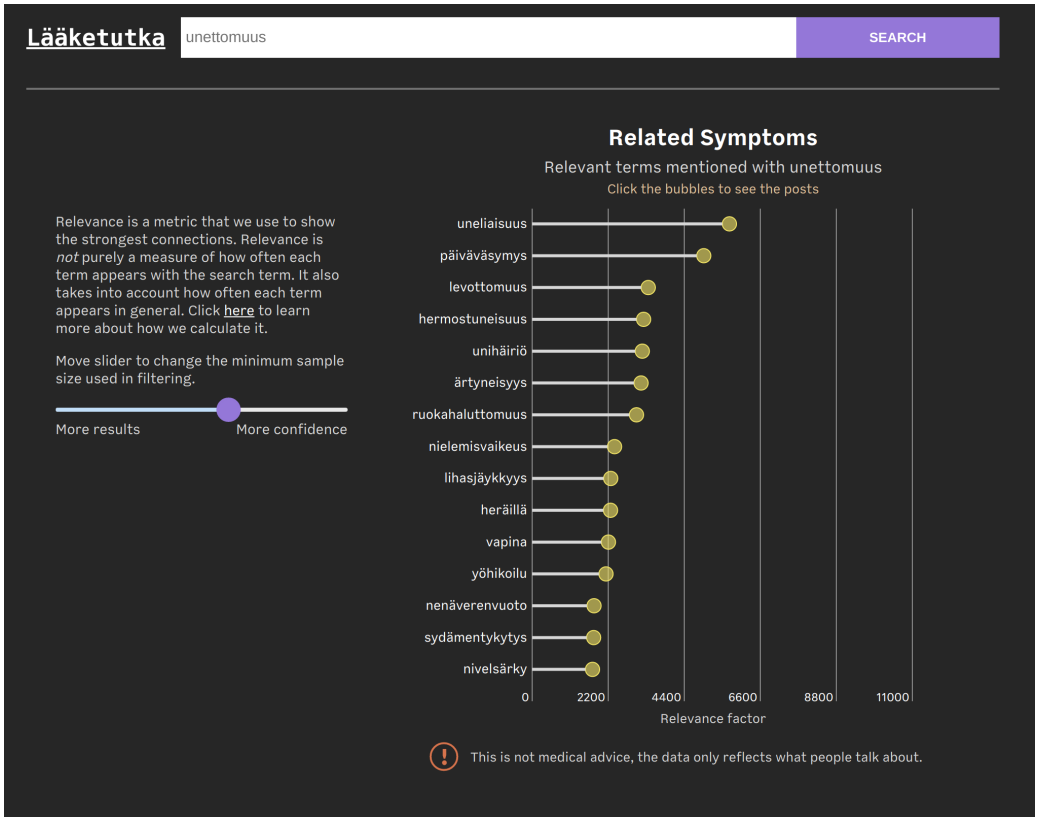

Fig. 1. Symptoms related to sleeplessness

offers visualization of most common dosages (Figure 3). Clicking on a dosage allows the user to read the related discussions. Dosages were identified based on a regular expression, and linked to the closest medicine concept within the same post.

4 What is Medicine Radar good for?

We developed Medicine Radar for facilitating research and public debate: it allows researchers with no technical skills to access a large social media data set. The goal of the tool is to give access to the discussion landscape, its patterns and propensities, in order to highlight how drugs are perceived and lives are shaped by the use of drugs in ways not evident to those who develop, research or administer said drugs as part of medical treatment protocols.

As a qualitative research tool, Medicine Radar provides support for answering a plethora of questions that concern everyday experiences of health and illness. We know from existing research that studies of pharmaceuticals tend to give

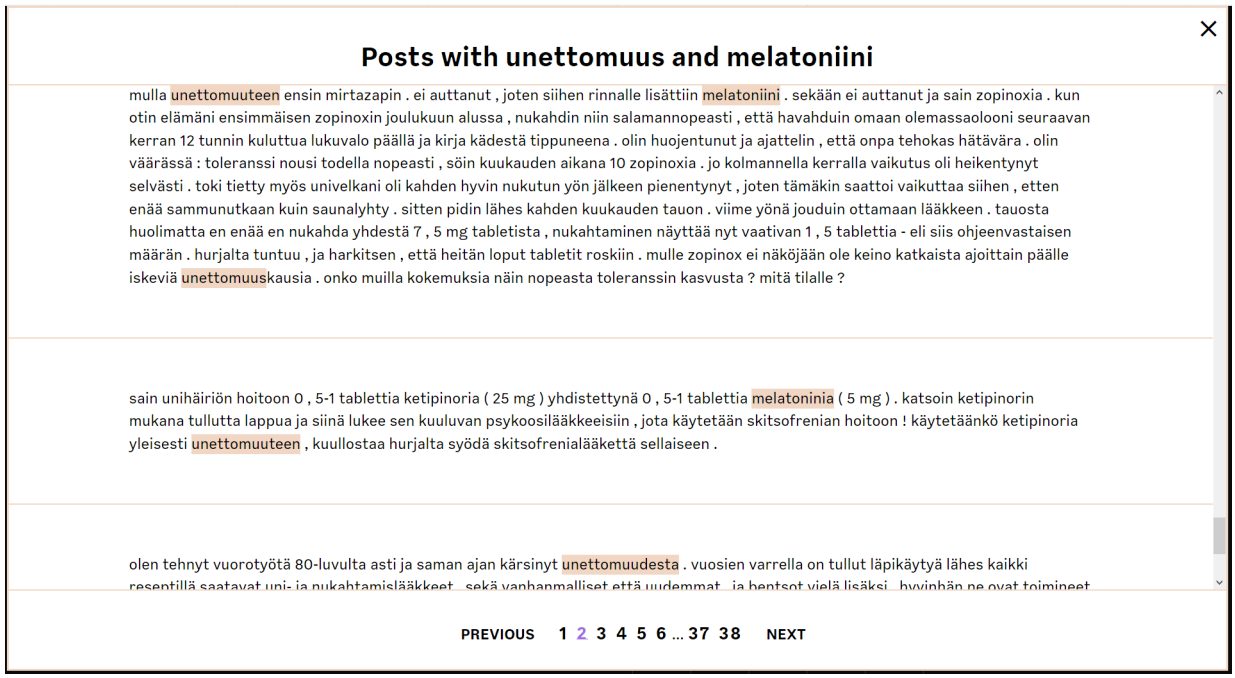

Fig. 2. A list of posts containing both ‘melatonin’ and ‘sleeplessness’.

a too flawless and straight-forward notion of the human-drug relationship. Patients don’t necessarily take their medicines according to the instructions given; the treatments are left in the middle, the doses are irregularly taken or not taken as prescribed. Medicines are borrowed and improvised with, formerly prescribed drugs are being re-used as the symptoms get worse [15] [16] [17] [18] [19]. In addition to shedding light to these questions, Medicine Radar can support the study of conversations around distinct medicines, symptoms, or experiences of some particular disease. Based on the feedback from researchers, medical professionals and students, the tool is seen as useful for exploring how people medicate themselves, offer peer-advice and engage in quack-doctoring. The discussions reveal a universe of personal health histories: painful personal experiences, long-lasting health problems and side effects of medication.

Medicine Radar can also be used in the education for medical doctors and pharmacists. It can help raise awareness of issues which are considered too intimate or controversial to be talked about with a pharmacist or doctor. As an additional example, concerning drug abuse and addictions, Medicine Radar can offer support for developing services in these fields. In service development, it is essential to understand the language and the experiences of intended users, and Medicine Radar offers first-hand material towards that goal.

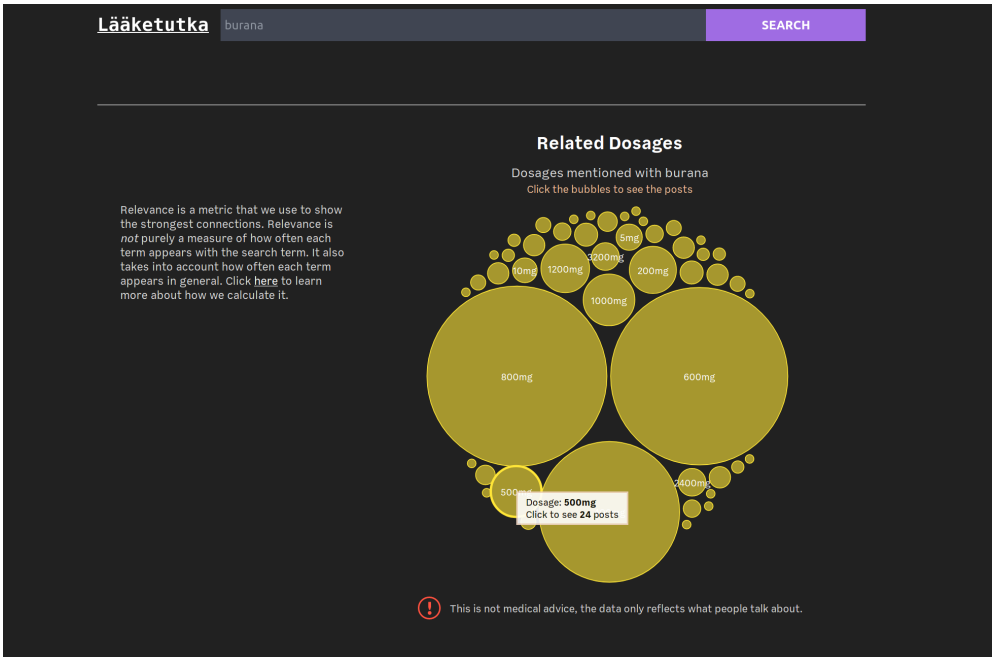

Fig. 3. Dosages most often associated with "Burana".

5 Acknowledgements

We thank Päivi Metsäniemi (Terveystalo) and Antti Ajanki (Futurice) for their generous advice. Futurice's Chilicorn Fund supported the research and development of Medicine Radar: we thank Mustafa Saifee for data visualisations, Maritere Vargas for website design, Christian Fricke for help on database optimization as well as Teemu Turunen for continued support. The Citizen Mindscapes initiative gratefully acknowledges the support from Finnish Academy (grant 292906), and the resources of the CSC and the Language Bank. Finally, we are grateful to Aller for opening the Suomi24-data for research purposes.

References

1. Creutz, M and Lagus, K. (2002).: Unsupervised discovery of morphemes. In Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02, pages 21-30, Philadelphia, PA, July 11.
2. Creutz, M. and Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. ACM Transactions on Speech and Language Processing, 4(1), January 2007.
3. De Choudhury, M., & De, S. (2014, June). Mental Health Discourse on Reddit: Self-Disclosure, Social Support, and Anonymity. In ICWSM - Eighth International AAAI Conference on Weblogs and Social Media.

4. Ginter, F. and Kanerva, J. (2014). Fast training of word2vec representations using n-gram corpora. SLTC 2014.
5. Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., Ojala, S., Salakoski, T. & Ginter, F. (2014). Building the essential resources for Finnish: the Turku Dependency Treebank. Language Resources and Evaluation, 48(3), 493-531.
6. Honkela, T., Pulkki, V., & Kohonen, T. (1995). Contextual relations of words in Grimm tales analyzed by sself-organizing map. In Proceedings of ICANN-95, international conference on artificial neural networks (Vol. 2,pp. 3-7).
7. Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.
8. Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., and Saarela, A. (2000) Self organization of a massive text document collection. Transactions on Neural Networks. Special Issue on Neural Networks for Data Mining and Knowledge Discovery, 11 (3), pp. 574-585. © 2000 IEEE.
9. Lagus, K., Honkela, T., Kaski, S., and Kohonen, T. (1996). Self-organizing maps of document collections: A new approach to interactive exploration. In Simoudis, E., Han, J., and Fayyad, U., eds., Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining, pp. 238-243. AAAI Press, Menlo Park, CA.
10. Lagus, K., Airola, A. and Creutz, M. (2002). Data analysis of conceptual similarities of Finnish verbs. In Proceedings of the CogSci 2002, the 24th annual meeting of the Cognitive Science Society, pages 566-571. Fairfax, Virginia, August 7-10, 2002.
11. Lagus, K. H., Pantzar, M., Ruckenstein, M. S. & Ylisiurua, M. J. (2016) Suomi24 - Muodonantoa aineistolle (Suomi24 - Giving shape to the data set). Valtiotieteellisen tiedekunnan julkaisuja 10, May 2016. Helsinki: Unigrafia. 44 pages.
12. Lagus, K. H., Pantzar, M., Ruckenstein, M. S. (To appear) Tunneaallot verkkokeskustelussa ja kulutustutkimuksessa (Emotional waves in online conversation and consumer research). Kulutustutkimus.Nyt (to appear).
13. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
14. Ritter, H., & Kohonen, T. (1989). Self-organizing semantic maps. Biological cybernetics, 61(4), 241-254.
15. Pound, Pandora, Nicky Britten, Myfanwy Morgan, Lucy Yardley, Catherine Pope, Gavin Daker-White and Rona Campbell. 2005. "Resisting Medicines: A Synthesis of Qualitative Studies of Medicine Taking." Social Science & Medicine 61 (1): 133–55.
16. Stevenson FA, Britten N, Barry CA, et al. (2003) Self-treatment and its discussion in medical consultations: how is medical pluralism managed in practice? Social Science & Medicine 57: 513-527.
17. Ylisiurua, Marjoriikka 2017: Aihemallinnuksen mahdollisuudet sosiaalisen median aineistojen jäsentämisessä – terveyskeskustelu Suomi24verkkopalstalla. Kulutustustkimus. Nyt (11) 2/2017.
18. Weiner, K and Will, C (2016) Use, non-use and resistance to pharmaceuticals, in Hyysalo, S., Jensen T. and Oudshoorn, N, (eds) The New Production of Users: Changing innovation collectives and involvement strategies. Routledge,Abingdon, Oxon. pp. 273-296.
19. Will, C and Weiner K. (2015). The Drugs Don't Sell: DIY Heart Health and the Over-the-Counter Statin Experience. Social Science & Medicine 131: 280–88.